International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 8, ISSN No. 2455-2143, Pages 94-98
Published Online June - July 2016 in IJEAST (http://www.ijeast.com)

FUSION BASED VIDEO TEXT DETECTION APPROACH

Pooja
Department of CSE
Dr B R Ambedkar NIT, Jalandhar, Pb., India

Renu Dhir
Department of CSE
Dr B R Ambedkar NIT, Jalandhar, Pb., India

***Abstract*— Text extraction from pictures has forever been a challenge for the researchers since long. Lots of Techniques are developed and experimented for similar work. Text detection from the videos is incredibly necessary and in demand currently for pertinency in varied areas viz. video categorization, retrieval, content analysis, annotation. This paper presents a fusion based approach to find text within the video frames. The text thought of is particularly the news ticker, that typically appear at the lower 1/4th height of the (news) video as discovered. The approach uses Laplacian of Gaussian for edge detection and Otsu for foreground detection. Further the outcomes of those two algorithms are fussed into one image by using principal component analysis technique. This fusion based extraction of text can further is to be used further by us to develop a replacement algorithmic rule to find scrolling text in videos. In this paper we tend to demonstrate the results of text extraction by conducting experiments on quite a variety of frames of a news video.**

Keywords— **Video image, frame extraction, text detection, OCR**

I. INTRODUCTION

Algorithms should be developed that may automatically extract linguistics data from video using content alone. Given an arbitrary video sequence, such algorithms would verify as abundant data as attainable, like genre (sitcom, movie, sports program, etc.), cinematography location characteristics (indoor or outside, time of day, weather conditions, etc.), identity of vital objects, identity of individuals (politicians, film stars, programme characters, etc.), and human action and interaction (running, laughing, talking, arguing, etc.). This wealth of data may well be accustomed better identify video sequences of interest to a user. Automatically extracting this information from unconstrained video is extremely difficult. Solving the underlying computer vision and AI issues will beyond any doubt occupy these research communities for several years. Additionally to the features mentioned above, text appearing during a video sequence will offer helpful linguistics information. Text occurring in video naturally provides clues to the video's content. Words have well-defined, unambiguous meanings. If the text in a video sequence can be extracted, it will provide natural, significant keywords indicating the video's content. Optical character recognition (OCR) of document pictures has been studied extensively for many years [1]. Technology has evolved to almost solve the document OCR problem. Recognition accuracy rates over 99 are currently possible. However, extraction of text from video presents distinctive challenges over OCR of document images. Document pictures are typically scanned at high resolutions of 300 dots per inch or higher. In distinction, video frames are usually digitized at much lower resolutions, usually 640x480 or 320x240 pixels for a complete frame. Additionally, lossy compression schemes are usually applied to digital video to keep storage necessities affordable. Video frames thus suffer from color bleeding, loss of contrast, block artifacts, and other noise that considerably will increase the problem of accurately extracting text. Several characteristics of the text in a document image are best-known a priori. As an example, the text color in a document is almost always black, and also the background is known to be uniform white. There is high distinction between the background color and also the text color. The orientation of the text is often assumed to be horizontal, or can simply be inferred by analyzing the structure of the document. In contrast, text in video can have whimsical and non-uniform stroke color. The background could also be non-uniform, complex, and ever-changing from frame to frame. The contrast between the background and foreground could also be low. Text size, location, and orientation are at liberty too. The temporal nature of video introduces a new dimension into the text extraction problem. Text in video sometimes persists for at least several seconds, to allow human viewers the required time to read it. Some text events stay unchanged throughout their lifetimes. Others, like film credits, move in a straightforward, rigid, linear fashion. Still others like scene text and stylized caption text move and alter in complicated ways. Text can grow or shrink, or character spacing can increase or decrease. Text color can change over time. Text can rotate and alter orientation. Text can morph from one font to a different. Text strings can break apart or join along. Special effects or a moving camera can cause dynamical text perspective.

The problem of text extraction from video is thus considerably tougher than the document image OCR problem. It is attainable to change the problem by creating a priori assumptions

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 8, ISSN No. 2455-2143, Pages 94-98
Published Online June - July 2016 in IJEAST (http://www.ijeast.com)

concerning the kind of video, or to extract solely certain kinds of text. However, in a general-purpose video indexing application, it is vital to be able to extract as much text as possible. Therefore text extraction systems must be applicable to all-purpose video data and must be ready to handle as many varieties of text as attainable.

This paper discusses the detection of unconstrained caption text from general purpose video. Especially, it addresses the detection of text present in videos that have not been worked upon largely in the literature so far. The main focus of this work is on extraction of news ticker text.

The remainder of the paper discusses the detection issues as in Section II, In Section II(A), framework of the fusion based method has been discussed. Section III extends this work to allow detection of text through the experimental setup. Finally, the results are conferred in Section IV.

II. TEXT DETECTION

A digital video is a sequence of still images, displayed quickly to provide the illusion of continuous motion. Locating text in video therefore begins with locating text in images. This chapter considers the problem of distinguishing text regions in images and video frames. The process of identifying text regions are often split into 2 sub problems: detection and localization. Within the detection step, general regions of the frame are classified as text or non-text. The scale and form of these regions differ from algorithm to algorithm.

For example, some algorithms classify 8X8 pixel blocks, whereas others classify individual scan lines. In the localization step, the results of detection are classified along to make one or a lot of text instances.

There are 2 varieties of text in videos:

Scene Text: the text that appears during a video scene is termed the scene text as an example text written on billboards, buildings, banners, shirts etc. it is also stated as graphic text.

Artificial text: The text that is superimposed by artificial means into a video by editing devices once the video is created is termed the artificial text. It appears on the screen at some specific position e.g. horizontally at the lower part of the screen. Some of the Examples are names, data concerning the video or translation of dialogues in video motion picture, subtitles during a video. It is conjointly named as superimposed text or caption text.

A novel dimension gets established by the temporal nature of video, into the text extraction problem. Text in video usually perseveres for more than a few seconds. Some text events remain unchanged during their lifetimes. Others, like movie credits, move in a straightforward, direct manner. Whereas ones, like scene text and stylized caption text, move and change in multifaceted behavior. Text can grow or get smaller, or space between the characters can increase or decrease. Text color can change over time. Text can spin and vary its orientation. Text can morph from one font to another. Text strings can break apart or join together. Special effects or a moving camera can cause changing perspective. Text detection involves locating regions in the frame. Text localization groups text regions identified by the detection stage into text instances. The localization algorithm yields a set of tight bounding boxes around all text cases.

Same text string in digital video, often spans tens or even hundreds frames so multiple frame information can be explored to enhance the textual image quality. There are two types of pixels in each text block: text pixel and background pixel. In each frame the intensities of text pixels and background pixels are often mixed and binarization will lead to correctness. If we observe the same text string over a long sequence however, we can see that the text pixels are relatively constant while background pixels keep changing. This is clearly true when motion exists in the background and text is independent. But even in cases where the back- ground is not moving, slight changes in intensity are often present due to noise. If text blocks can be registered and averaged to generate a new text block, then the text pixels in the new text block will remain unchanged while the background tends to be smoothed, making the text separation easier.

A. Text Extraction Framework –

First of all the Punjabi videos of News are downloaded from the internet and are stored in a folder for making the raw database. These videos are then input to the frame extraction block. In this block Matlab code is used to extract out the frames from the videos. All the frames extracted are stored in the separate folder for creating the dataset of the frames. From the frame dataset, frames are fed to the next two blocks, the LoG Based Edge detection and Otsu based Foreground extraction blocks. The last step in the framework is to fuse the outcome of these blocks into a single image.

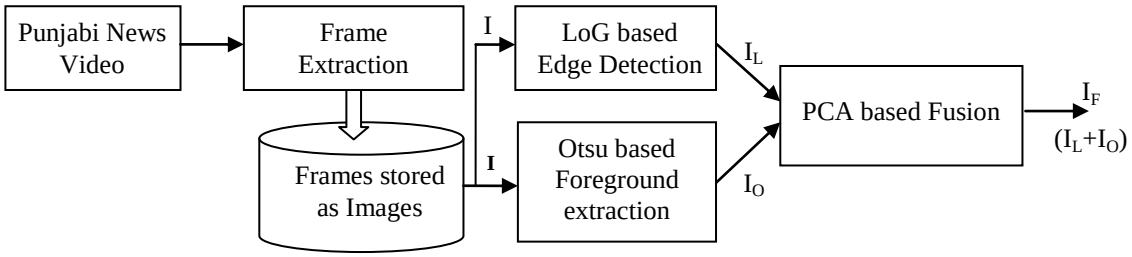


Fig. 1. Fusion based text extraction

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 8, ISSN No. 2455-2143, Pages 94-98
Published Online June - July 2016 in IJEAST (http://www.ijeast.com)

III. EXPERIMENTAL SETUP

For carrying out the experiment, foremost requirement is the availability of the video containing the news ticker. As this work is done for the detection of Punjabi (Gurmukhi) text, we need to have the videos which contain text in Punjabi language only. Internet has been the paramount source of information these days. So, internet was considered for obtaining news videos for experimenting with the methodology proposed.

Further we need to extract the frames from each of these videos for performing the experiments. For the extraction of frames Matlab is used as it is easy and computation is fast. For each available video a separate folder is created to store the frames. All the frames are stored sequentially. To refer to every single frame, we named them using abbreviation 'ViFj'. Here i represent the number of video and it ranges from 1-15, while j represents the number of frame it may range from 1-∞, (depending upon the number of frames a video may contain).

Next is how to detect the text present in the frames? For this purpose, two different algorithms are used that too on the same frame image taken from the frame repository. The two algorithms are: - Laplacian of Gaussian, LoG, a very powerful filter to extract edges i.e. Text. And the other is Otsu algorithm, another best algorithm based on threshold to detect even the squeeze and broken edges.

Both the outcomes are then blending as one into a solo image. Fusion is done to make sure that no text part is left undetected. The fused image will be full of text edges only leaving behind the unwanted part of the video (as far as this research is concerned, text the only important object in the video).

For the fusion of two images Principle Component Analysis is chosen, just to make it less complex and to get better results. PCA is the straightforward and mainly helpful in the genuine eigenvector-based multivariate investigations, since its operation is to uncover the interior arrangement of information neutrally. On the off chance that a multivariate Dataset is imagined as an arrangement of directions in a hoisted dimensional information space, PCA supplies the client with a 2-Dimensional picture, a sad remnant of this subject when prospect from its most enlightening perspective. This dimensionally-lessened picture of the information is the appointment figure of the first two foremost tomahawks of the information, which when united with metadata can quickly expose the main factors underlying the structure of data. Principal component analysis (PCA) is a vector space Transform often used to ease multidimensional data sets to lower dimensions for analysis.

Image fusion process via PCA can be described as follow:

1. From the input image matrices create the column Vectors.
2. Calculate the covariance matrix of two column vectors twisted before.
3. Determine the Eigen value and Eigen vector of the Covariance matrix.
4. The column vector parallel to the superior Eigen value is normalized by dividing each element with mean of Eigen vector.
5. Stabilize Eigen vector value act as the weight Values which are respectively multiplied with each Pixel of the input images.
6. The fused image matrix will be computation of the two scaled matrices.

It is a numerical tool which changes correlated variables into uncorrelated ones. Another set of axes are produced as a result which are orthogonal in nature. The first chief component is taken with the most extreme fluctuation. The second main component is constrained in the subspace opposite to the first computer and it goes on so [9]. This algorithm replaces the spatial segment of the multi-spectral image with that of the panchromatic image. It permits the spatial subtle elements of the panchromatic picture to be consolidated into the multi-spectral image. The principal component changes the resample paint bands of the multi-spectral image to the same resolution as the panchromatic image. The histogram that is matched to the first principal component accounts to be the panchromatic image formed [10]. This is done in order to remove the spectral differences between the two images, which happened because of different types of sensors or distinctive procurement of dates and edges. The histogram matched panchromatic imagery replaces the primary essential component of the multi-spectral image. By computing the converse of principal component transformation, the new consolidated multi-spectral imagery is acquired. It is worth mentioning here that all these algorithms are developed in Dotnet. And to keep a check on the code, (if it is being written/executing rightly and if the results are as expected in the form of images?), Matlab version of the same was also developed Parallely.

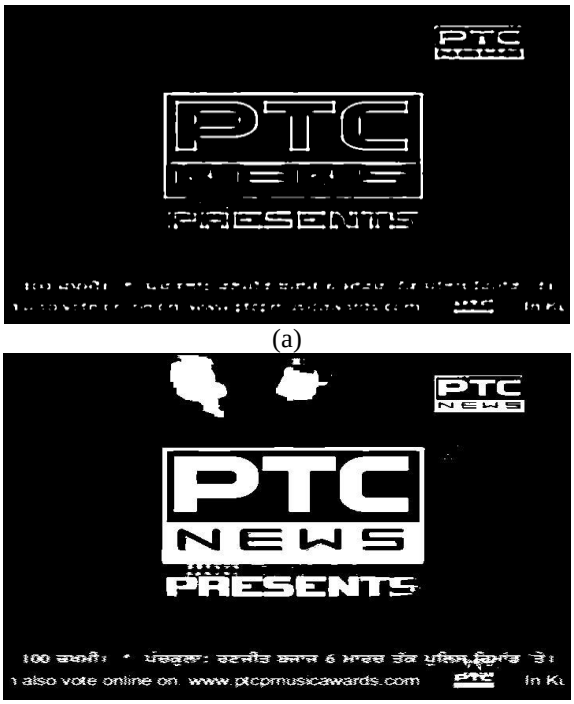
IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

This paper presents the pictorial results of the experiments conducted on frames of one news video and the tabular results of the experiments conducted on the whole dataset of frames extracted from all videos collected. Figure 2 displays two of the frames extracted out of the news video, as a sample. As explained in the previous section of this paper, the frames are extracted from the videos by using MatLab.

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 8, ISSN No. 2455-2143, Pages 94-98
Published Online June - July 2016 in IJEAST (http://www.ijeast.com)

Fig. 2. Successive two Frames extracted from videos

These frames are then Parallely processed through two different algorithms viz. LoG and Otsu. The following figure (Figure 3) is showing the outcome of these algorithms respectively of the first two frames consecutively.



(a)

(b)

Fig. 3. (a) LoG Image, (b) Otsu Image

The experiments conducted on all of the frames and results are reported in the tables for each video. The duration of the videos from which the frames are extracted is kept 3-5 minutes, as the number of frames range from 2000 to 4000 for such a small duration only. Figure 4 exhibits the fused image.

Fig. 4. Fused Image (LoG n Otsu output of frame 1)

V. CONCLUSION

The output images are better than the existing methods used to detect the text from the images. Further the results can be extended to extract the text from the fused images and input to the OCR for text recognition. Detection is the base of any text recognition process. If the detection has been done properly, any best algorithm would produce result in a better way.

VI. REFERENCE

[1] T. Sato, T. Kanade, E. K. Hughes, and M. A. Smith. "Video OCR for Digital News Archive", in *Proceedings of IEEE Workshop on Content based Access of Image and Video Databases*, pp. 52-60, 1998.

[2] G. Nagy, "Twenty Years of Document Image Analysis in PAMI", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 1, pp. 38-62, 2000.

[3] Li Huping, Doermann D and Kia O, "Automatic Text Detection and Tracking in Digital Video", *IEEE Transactions on Image Processing*, Vol. 9, No. 1, ISSN 147-156, 2000.

[4] Datong Chen, Jean-Marc Odobez, Herv/e Bourlard, "Text detection and recognition in images and video frames", *The Journal of the Pattern Recognition Society*, Vol. 37, pp. 595 – 608, 2004.

[5] M. Anthimopoulos, B. Gatos, I. Pratikakis, S. J. Perantonis, "Detecting Text in Video Frames", in *Proceedings of The Fourth IASTED International Conference on Signal Processing, Pattern Recognition, and Applications*, pp. 39-44, 2007.

[6] Zhiyi Zhang, Lianwen Jin, Kai Ding, Xue Gao, "Character-SIFT: a novel feature for offline handwritten Chinese character recognition", in *Proceedings of IEEE*

10th International Conference on Document Analysis and Recognition, pp. 763-767, 2009.

[7] Shivakumar P, "A Laplacian Approach to Multi-oriented Text Detection in Video", *IEEE Transactions on* pattern analysis and machine intelligence, Vol. 33, No. 2, pp. 412-419, 2011.

[8] Nabin, "A New Method for Arbitrarily-Oriented Text Detection in Video", in Proceedings of 10th IAPR International Workshop on Document Analysis Systems (DAS) ISBN 978-1-4673-0868-7, pp. 74-78, 2012.

[9] V. P. S. Naidu and J. R. Raol, "Pixel-level Image Fusion using Wavelets and Principal Component Analysis", Vol. 58, No. 3, pp. 338–352, 2008.

[10] A. Pradesh, "Edge Preserving Satellite Image Enhancement Using DWT-PCA Based Fusion and Morphological Gradient", in *Proceedings of IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT),* pp. 1–5, 2015.